

Bootstrapping instance-based ontology matching via unsupervised generation of training samples

ABSTRACT

Training set is the key role player that can improve the performance of any classification task. Different techniques and methods are being applied to generate training set depending on its area of application. Researchers in data science and semantic web community use different kind of training sets generated to improve the performances of classifications and information retrieval capability. Operational Training Set Generator (TSG) should always solve a minimum of two issues; (1) it must address the computational cost in producing a reasonable outcome, thereby reducing the computational cost in the whole system. The runtime of TSG is near linear as in blocking approach and (2) it must produce the qualitative training sets. We use LogTfIdf as the cosine similarity function of two given vectors to produce Bag of Words (BoW); the tokenizer is developed to specially take care of delimiters that often come across URIs and other RDF essentials. We evaluated our UTSG on nine cross-domain benchmark ontologies publically available in OAEI website. The results obtained shows that our UTSG outperforms the two baseline TSGs previously developed to address similar problem.

Keyword: Semantic web; Link open data; Semantic heterogeneity; Ontology matching; Instance-based matching; Training set